Name, SURNAME and ID ⇒ SOLUTIONS

Middle East Technical University
Department of Computer Engineering

CENG 465

Introduction to Bioinformatics

Spring '2006-2007

Midterm Exam II

- **Duration: 90** minutes.
- **Exam:**
  - This is a **closed book, closed notes** exam. The use of any reference material is strictly forbidden.
  - No attempts of cheating will be tolerated. In case such attempts are observed, the students who took part in the act will be prosecuted. The legal code states that students who are found guilty of cheating shall be expelled from the university for **a minimum of one semester**!
- **About the exam questions:**
  - The points assigned for each question are shown in parenthesis next to the question.
  - For *True-False* type questions, put your results in the boxes provided.
- **This exam consists of 7 pages including this page. Check that you have them all!**
- **GOOD LUCK !**

Question 1	20
Question 2	20
Question 3	30
Question 4	30
Total ⇒	100

1 **(20 pts)** 20

For the following 10 statements, indicate whether the statement is *true* or *false* by marking the corresponding box with **T** or **F**, respectively (2 points each).

i. The UPGMA algorithm always produces binary trees. T

ii. All symmetric matrices are *additive*. F

iii. The path distance between two leaf nodes in a phylogenetic tree is equal to the sum of the distances from each node to the root node. F

iv. Two amino acids that are far apart in the primary structure of a protein can be close to each other in the tertiary structure of that protein. T

v. The secondary structure of a protein can be described by a character string, i.e., a linear sequence of letters. T

vi. In general, secondary structure prediction algorithms predict *alpha helices* more accurately compared to *beta sheets*. T

vii. In order to be able to predict the tertiary structure of a protein sequence q using *threading*, q should have some sequence similarity to a protein with known tertiary structure. F

viii. In protein structure comparison, the *root mean squared distance* between two aligned protein structures is computed after the aligned amino acids are *superimposed*. T

ix. DALI algorithm for protein structural alignment aligns intra-atomic distance matrices of protein structures. T F is also accepted. (intra-atomic vs. intra-molecular)

x. STRUCTAL algorithm for protein structure comparison is an iterative dynamic programming algorithm. BLOSUM62 matrix is used in the dynamic programming component of the algorithm to score amino acid pairings. F

2 **(20 pts)**

20

Prove that the *sum of pairs score* of a multiple sequence alignment is always less then or equal to sum of the alignment scores of Needleman-Wunsch dynamic programming alignments of all pairs of sequences. Assume that same scoring matrices and gap penalties are used and a *gap-gap* alignment in the multiple sequence alignment has 0 (zero) score.

You may provide an informal proof with a couple of sentences. The statement above is not specific to a certain multiple sequence alignment algorithm; your proof should hold for all multiple sequence algorithms.

The Needleman-Wunsch algorithm always produces the optimum global pairswise alignment between two sequences. The induced pairwise alignments by the MSA are occasionally non-optimum pairwise alignments. In the best case that the induced alignments are optimum alignments the Sum of Pairs score can be equal to the pairwise NW scores, but they can never exceed.

3 (30 pts)

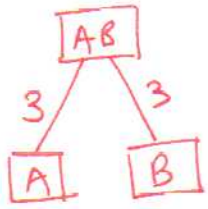
30

Given the following additive distance matrix between 5 protein sequences.

	A	B	C	D	E
A	0	6	13	15	11
B		0	11	13	9
C			0	12	8
D				0	10
E					0

(a)(20 pts) Use the UPGMA algorithm to construct a phylogenetic tree of the sequences represented by the distance matrix. Show intermediate matrices and subtrees in your construction.

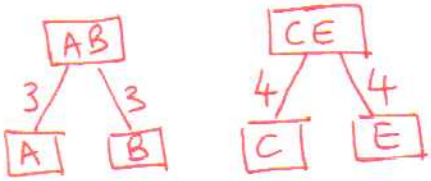
Step 1: Combine A and B



Updated distance matrix:

	AB	C	D	E
AB	0	12	14	10
C		0	12	8
D			0	10
E				0

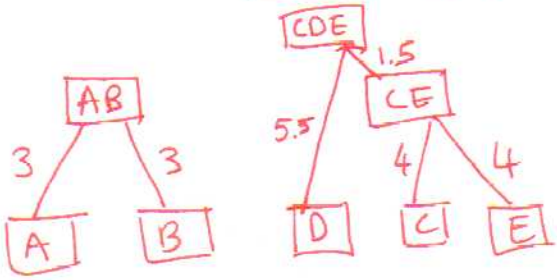
Step 2: Combine C and E



Updated distance matrix:

	AB	CE	D
AB	0	11	14
CE		0	11
D			0

Step 3: Combine either AB and CE
or CE and D
I choose to combine CE and D

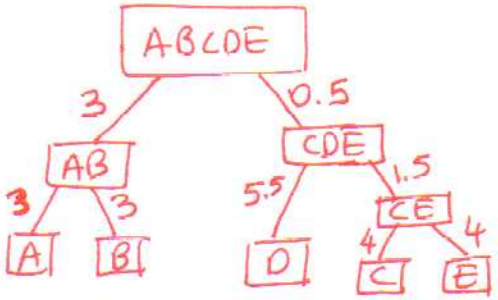


Updated distance matrix: $d_{AB,CDE} = \frac{2 \cdot 11 + 14}{3} = \frac{36}{3} = 12$

	AB	CDE
AB	0	12
CDE		0

Step 4 (Final Step):

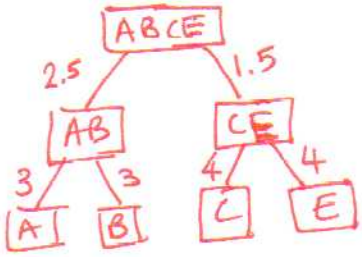
Combine AB and CDE



Final Tree

Alternative Step 3:

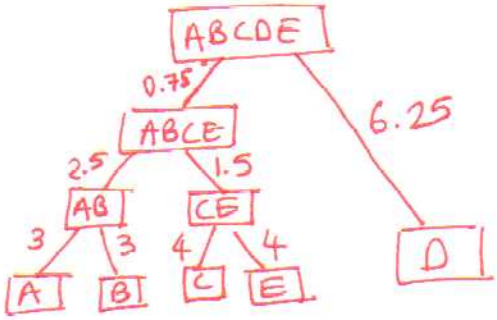
Combine AB and CE



Updated distance matrix:

	ABCE	D
ABCE	0	12.5
D		0

Alternative final tree:



(b)(10 pts) Write down the *path distance matrix* based on the tree you have constructed in step (a). Is the path distance matrix equal to the original distance matrix given in question? If not, what additional property does the distance matrix need to satisfy so that the path distance matrix is equal to the original distance matrix?

path distance matrix d_{ij}

	A	B	C	D	E
A	0	6	12	12	12
B		0	12	12	12
C			0	11	8
D				0	11
E					0

path distance matrix for the alternative tree:

	A	B	C	D	E
A	0	6	11	12.5	11
B		0	11	12.5	11
C			0	12.5	8
D				0	12.5
E					0

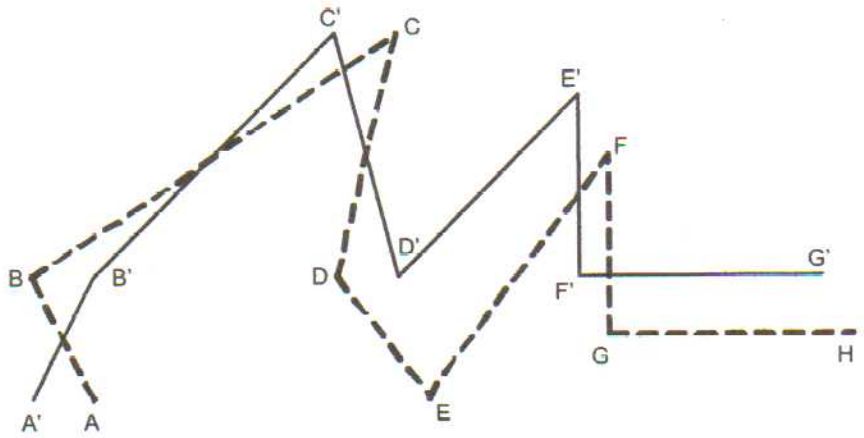
The path distance matrix is not equal to the original distance matrix in both solutions.

The original matrix should have been an ultrametric matrix so that the path distance matrix is equal to the original distance matrix.

4 **(30 pts)**

30

Given the two following aligned and superimposed protein structures in 2D.



The first protein P_1 is shown by dashed lines and specified by the following carbon alpha atoms with respective coordinates:
A(0,0)~B(-4,8)~C(20,24)~D(16,8)~E(22,0)~F(34,16)~G(34,4)~H(50,4)

The second protein P_2 is shown by solid lines and specified by the following carbon alpha atoms with respective coordinates:
A'(-4,0)~B'(0,8)~C'(16,24)~D'(20,8)~E'(32,20)~F'(32,8)~G'(48,8)

The alignment shown in the figure is obtained by aligning ABCD with A'B'C'D', i.e., the structural alignment of P_1 with P_2 given in the figure above matches substructure ABCD with A'B'C'D'.

(a)(10 pts) Compute the RMSD for the alignment given above.

The alignment given above aligns ABCD with A'B'C'D'. Therefore, we need to consider the distances between those points when computing the RMSD for that alignment.

$$RMSD = \sqrt{\frac{d_{AA'}^2 + d_{BB'}^2 + d_{CC'}^2 + d_{DD'}^2}{4}} = \sqrt{\frac{4^2 + 4^2 + 4^2 + 4^2}{4}}$$

$$= \sqrt{16} = 4 \text{ units}$$

(b)(10 pts) Is there a better alignment of same length with a better RMSD? If so, show the alignment by specifying matched amino acids and compute the new RMSD.

Yes, there is a better alignment of length 4 which aligns:

D' E' F' G'
E F G H

To compute new RMSD we need to superimpose the proteins based on the new alignment with the translation $T(-2,+4)$ applied to the dashed structure. New RMSD then becomes

$$RMSD = \sqrt{\frac{d_{D'E} + d_{E'F} + d_{F'G} + d_{G'H}}{4}} = \sqrt{\frac{4^2+0^2+0^2+0^2}{4}} = 2 \text{ units}$$

(c)(10 pts) Describe why the structural alignment algorithm STRUCTAL preserves sequence order in the produced structural alignments. How does DALI produce alignments in which sequence order is not preserved?

Because STRUCTAL uses dynamic programming to align structures. DP preserves sequence order (the 2D score matrix is created respecting the sequence order)

DALI can combinatorically combine similar submatrices from different locations in different order; therefore does not respect sequence order.